The mathematical foundations of deep learning: from rating impossibility to practical existence theorems

"Get ready to delve into the mind-bending intersection of mathematics and artificial intelligence, where elegant equations and concepts lay the foundation for the miraculous advancements in deep learning that are transforming our world today."

Simone Brugiapaglia

http://simonebrugiapaglia.ca

Concordia University

March 30, 2023

Biomedical Imaging for Healthy Aging Lab Seminar, Concordia

Biomedical Imaging for Healthy Aging Laboratory
Bhealthyage

Acknowledgements

Collaborators

Concordia
Matthew Liu

Florida State University
Nick Dexter

Simon Fraser University
Ben Adcock
Sebastian Moraga
Paul Tupper

University of Texas at Austin
Clayton Webster

Funding

NSERC
CRSNG

Université Concordia University

Centre de Recherches Mathématiques

Pacific Institute for the Mathematical Sciences

Who is this quote from?

> "Get ready to delve into the mind-bending intersection of mathematics and artificial intelligence, where elegant equations and concepts lay the foundation for the miraculous advancements in deep learning that are transforming our world today."

ChatGPT!

S

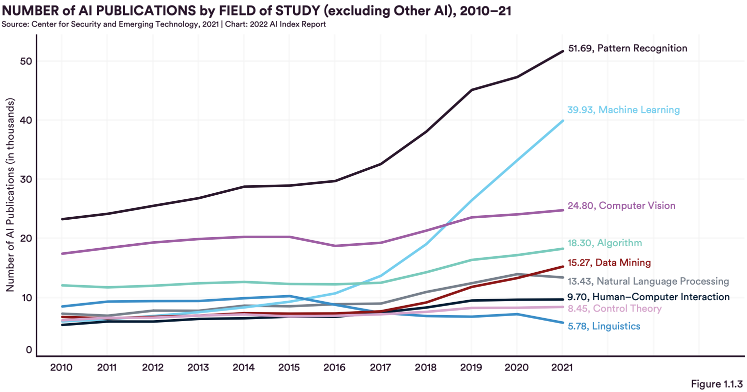
Write a brief opening sentence for a seminar on the mathematical foundations of deep learning. Make it exciting and mind bending

"Get ready to delve into the mind-bending intersection of mathematics and artificial intelligence, where elegant equations and concepts lay the foundation for the miraculous advancements in deep learning that are transforming our world today."

Regenerate response

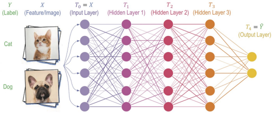
ChatGPT Jan 30 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

We live in an "AI golden age"

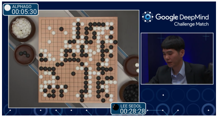


Source: 2022 AI Index Report (Stanford)
https://aiindex.stanford.edu/report/

The impact of deep learning



[© MIT-IBM Watson AI Lab]



[© DeepMind]

[© Tesla]

A pivotal role in the current AI revolution is played by **deep learning**.

Impactful applications include:

- AlphaGo project by DeepMind
- Speech synthesis in Apple Siri
- Speech recognition in the conversational engine of Amazon Alexa
- Netflix's recommender system
- Computer vision in Tesla's autopilot
- Conversational engine ChatGPT

The other side of the coin...

ANALYSIS | ARTIFICIAL INTELLIGENCE

2021's Top Stories About AI Spoiler: A lot of them talked about what's wrong with machine learning today

BY ELIZA STRICKLAND

27 DEC 2021 | 4 MIN READ

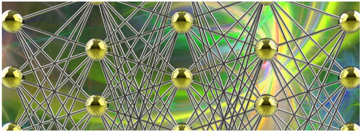


ILLUSTRATION BY EDGAR BÅK

DEEP TROUBLE FOR DEEP LEARNING

ARTIFICIAL-INTELLIGENCE RESEARCHERS ARE TRYING TO FIX THE FLAWS OF NEURAL NETWORKS.

BY DOUGLAS HEAVEN

A self-driving car approaches a stop sign, but instead of slowing down, it accelerates into the busy intersection. An accident report later reveals that four small rectangles had been stuck to the face of the sign. These fooled the car's onboard artificial intelligence (AI) into misreading the word 'stop' as 'speed limit 45'.

Such an event hasn't actually happened, but the potential for sabotaging AI is very real. Researchers have already demonstrated how to fool an AI system into misreading a stop sign, by carefully positioning stickers on it[1]. They have deceived facial-recognition systems by sticking a printed pattern on glasses or hats. And they have tricked speech-recognition systems into hearing phantom phrases by inserting patterns of white noise in the audio.

These are just some examples of how easy it is to break the leading pattern-recognition technology in AI, known as deep neural networks (DNNs). These have proved incredibly successful at correctly classifying all kinds of input, including images, speech and data on consumer preferences. They are part of daily life, running

10 OCTOBER 2019 | VOL 574 | NATURE | 163

POLICY FORUM

MACHINE LEARNING

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

By **Samuel G. Finlayson[1], John D. Bowers[2], Joichi Ito[3], Jonathan L. Zittrain[2], Andrew L. Beam[4], Isaac S. Kohane[1]**

With public and academic attention increasingly focused on the new machine learning into regulatory decisions by way of computational surrogate end points and so-called "in silico clinical trials."

Under the United States' health care model, some of the most direct impacts of machine-learning algorithms come in the context of

and across a w
ing images, au
Cutting-edge
erally use opti
data manipula
model. As a p
cal domain, w
ful adversarial
accurate medi
top figure prov
of these attack
commoditized
the left, an ima
which is corre
confidence of
what appears
fact a carefully
"adversarial no
to have maxir

Proving Existence Is Not Enough: Mathematical Paradoxes Unravel the Limits of Neural Networks in Artificial Intelligence

By Vegard Antun, Matthew J. Colbrook, and Anders C. Hansen

The impact of deep learning (DL), neural networks (NNs), and artificial intelligence (AI) over the last decade has been profound. Advances in computer vision and natural language processing have yielded smart speakers in our homes, driving assistance in our cars, and automated diagnoses in medicine. AI has also rapidly entered scientific computing. However, overwhelming amounts of empirical evidence [3, 8] suggest that modern AI is often non-robust (unstable), may generate hallucinations, and can produce nonsensical output with high levels of predic-

Our main result reveals a serious issue for certain problems; while stable and accurate NNs may provably exist, no training algorithm can obtain them (see Figure 2, on page 4). As such, existence theorems on approximation qualities of NNs (e.g., universal approximation) represent only the first step towards a complete understanding of modern AI. Sometimes they even provide overly optimistic estimates of possible NN achievements.

The Limits of AI: Smale's 18th Problem

The strong optimism that

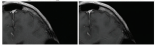
results about the feasible achievements of mathematics and digital computers.

A similar program on the boundaries of AI is necessary. Stephen Smale already suggested such a program in the 18th problem on his list of mathematical problems for the 21st century: *What are the limits of AI?* [11].

See **Mathematical Paradoxes** *on page 4*

Hallucinations in image reconstruction

Original image | AI reconstruction



The need for mathematical foundations

[Smale, 1998][1]

STEVE SMALE

Mathematical Problems for the Next Century[1]

V. I. Arnold, on behalf of the International Mathematical Union, has written to a number of mathematicians with a suggestion that they describe some great problems for the next century. This report is my response.

Arnold's invitation is inspired in part by Hilbert's list of 1900 (see, e.g., [Browder, 1976]) and I have used that list to help design this essay.

I have listed 18 problems, chosen with these criteria:

1. Simple statement. Also preferably mathematically precise.
2. Personal acquaintance with the problem. I have not

Problem 1: The Riemann Hypothesis

Of the zeros of the Riemann zeta function, defined by analytic continuation from

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \qquad \mathrm{Re}(s) > 1,$$

are those which are in the critical strip $0 \leq \mathrm{Re}(s) \leq 1$ *all on the line* $\mathrm{Re}(s) = \frac{1}{2}$?

Problem 18: Limits of Intelligence

What are the limits of intelligence, both artificial and human?

Penrose (1991) attempts to show some limitations of artificial intelligence. His argumentation brings in the interesting question whether the Mandelbrot set is decidable (dealt with in [Blum and Smale, 1993]) and implications of the Gödel incompleteness theorem.

However, a broader study is called for, one which involves deeper models of the brain, and of the computer, in a search of what artificial and human intelligence have in common, and how they differ. I would look in a direction where learning, problem-solving, and game theory play a substantial role, together with the mathematics of real numbers, approximations, probability, and geometry.

I hope to expand on these thoughts on another occasion.

[1] Written in reply to a request from Vladimir Arnold, then vice-president of the International Mathematical Union, who asked several mathematicians to propose a list of problems for the 21st century, inspired by Hilbert's list for the 20th century.

This talk

Our focus

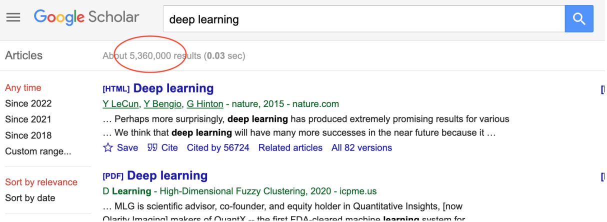
- Understanding the potential and limitations of deep learning through a rigorous mathematical approach

Two case studies

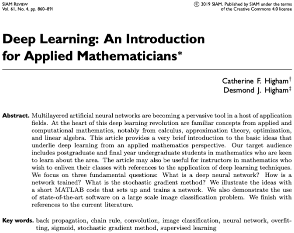
I. Rating impossibility theorems in identity effect classification

II. Practical existence theorems in high-dimensional approximation

Getting orientated amidst the DL literature “jungle”



Introductory paper:



SIAM Review
Vol. 61, No. 4, pp. 860–891

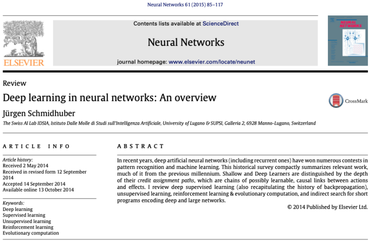
Deep Learning: An Introduction for Applied Mathematicians*

Catherine F. Higham†
Desmond J. Higham‡

Abstract. Multilayered artificial neural networks are becoming a pervasive tool in a host of application fields. At the heart of this deep learning revolution are familiar concepts from applied and computational mathematics, notably from calculus, approximation theory, optimization, and linear algebra. This article provides a very brief introduction to the basic ideas that underlie deep learning from an applied mathematics perspective. Our target audience includes postgraduate and final year undergraduate students in mathematics who are keen to learn about the area. The article may also be useful for instructors in mathematics who wish to enliven their classes with references to the application of deep learning techniques. We focus on three fundamental questions: What is a deep neural network? How is a network trained? What is the stochastic gradient method? We illustrate the ideas with a short MATLAB code that sets up and trains a network. We also demonstrate the use of state-of-the-art software on a large scale image classification problem. We finish with references to the current literature.

Key words. back propagation, chain rule, convolution, image classification, neural network, overfitting, sigmoid, stochastic gradient method, supervised learning

History:



Neural Networks 61 (2015) 85–117

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

ELSEVIER

Review

Deep learning in neural networks: An overview

Jürgen Schmidhuber

The Swiss AI Lab IDSIA, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, University of Lugano & SUPSI, Galleria 2, 6928 Manno-Lugano, Switzerland

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form 12 September 2014
Accepted 14 September 2014
Available online 13 October 2014

Keywords:
Deep learning
Supervised learning
Unsupervised learning
Reinforcement learning
Evolutionary computation

ABSTRACT

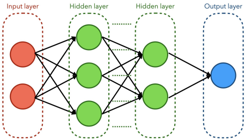
In recent years, deep artificial neural networks (including recurrent ones) have won numerous contests in pattern recognition and machine learning. This historical survey compactly summarizes relevant work, much of it from the previous millennium. Shallow and Deep Learners are distinguished by the depth of their *credit assignment paths*, which are chains of possibly learnable, causal links between actions and effects. I review deep supervised learning (also recapitulating the history of backpropagation), unsupervised learning, reinforcement learning & evolutionary computation, and indirect search for short programs encoding deep and large networks.

Deep neural networks (DNNs) in a nutshell

A (feedforward) Deep Neural Network (DNN) is a function approximator

$$\underbrace{x}_{\text{input}} \mapsto \underbrace{\sigma(\mathcal{A}_0(x))}_{=:h_1 \text{ hidden layer}} \mapsto \underbrace{\sigma(\mathcal{A}_1(h_1))}_{=:h_2 \text{ hidden layer}} \mapsto \cdots \mapsto \underbrace{\mathcal{A}_D(h_D)}_{\text{output}} = \Phi(x)$$

where the activation is, e.g., $\sigma(x) = \mathrm{ReLU}(x) = \max\{x, 0\}$ or $\sigma(x) = \tanh(x)$, and $\mathcal{A}_k$ are affine maps, i.e. $\mathcal{A}_k(x) = W_k x + b_k$.



[Image courtesy of Fahmi Nurfikri, towardsdatascience.com]

Architecture: Size of input, hidden, and output layers and choice of σ.

Depth: Number of hidden layers, D.

Trainable parameters: Define $\Theta = (W_k, b_k)_{k=0}^{D} \in \mathbb{R}^T$. Then, $\Phi = \Phi_\Theta$

Deep learning (DL) in a nutshell

Training: Given a dataset $\{(x_i, y_i)\}_{i=1}^m$, minimize a (regularized) loss:

$$\hat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^T} \underbrace{F((\Phi_\Theta(x_i), y_i)_{i=1}^m)}_{\text{loss function}} + \lambda \underbrace{R(\Theta)}_{\text{regularizer}}, \quad \lambda \geq 0$$

Examples:

- $F((\Phi_\Theta(x_i), y_i)_{i=1}^m) = \sum_{i=1}^m \ell(\Phi_\Theta(x_i), y_i)$, with

$$\ell(\phi, y) = \begin{cases} |\phi - y|^2 & \text{(least squares)} \\ -y\log(\phi) - (1-y)\log(1-\phi) & \text{(cross entropy)} \end{cases}$$

- $R(\Theta) = \|\Theta\|_p$, where $p = 1, 2$

Optimize via, e.g., **Stochastic Gradient Descent (SGD)**:

- define (random) partition $\{1, \ldots, m\} = B_0 \sqcup \ldots \sqcup B_{K-1}$ and $G_{B_j}(\Theta) = F((\Phi_\Theta(x_i), y_i)_{i \in B_j}) + \lambda R(\Theta)$,
- compute $\Theta_{j+1} = \Theta_j - \alpha_j \nabla_\Theta G_{B_{j \bmod K}}(\Theta_j)$, with $\alpha_j > 0$, and Θ_0 randomly initialized, until stopping criterion is met.

I. Rating impossibility theorems

Identity effects

Suppose you are told the following words are good:

AA GG LL MM

But that the following words are bad:

AG LM GL MA

Are the following words good or bad?

YY YZ

Identity effects

Suppose you are told the following words are good:

AA GG LL MM

But that the following words are bad:

AG LM GL MA

Are the following words good or bad?

YY YZ

Identity effect: well formedness depends on two substructures being identical.

Humans can easily generalize this type of task outside the training set (which did not contain Y, nor Z).

Can machine learning algorithms do the same?

Identity effects in cognitive science

Original motivation: linguistics [Benua, 1995; Gallagher, 2013].

- understanding whether a sentence is grammatical (syntax)
- or whether a word consisting of a string of phonemes is a possible word of a language (phonology).

[Marcus, 1999] shows that 7-month-old infants can generalize this type of rules, whereas neural networks cannot. This generated a heated debate.

Does generalization in infant learning implicate abstract algebra-like rules?

James L. McClelland and David C. Plaut

Connectionism: with or without rules?

Response to J.L. McClelland and D.C. Plaut (1999)

Gary F. Marcus

See also: *Boucher, V. (2020). Debate : Yoshua Bengio and Gary Marcus: The best way forward for AI.* `http://montrealartificialintelligence.com/aidebate/`.

Notation

$\mathcal{X}$ set of admissible inputs x

r rating (in $\mathbb{R}$)

Example:

$\mathcal{X}$ (set of all two-letter words, AA, LM, YY, ...)

$r \in [0, 1]$ (probability of being identical pair)

$\mathcal{D}$ training data set
Example:
$\mathcal{D} = \{(\text{AA}, 1), (\text{GG}, 1), (\text{LL}, 1), (\text{AG}, 0), (\text{LM}, 0), (\text{GL}, 0)\}$

$\mathcal{L}$ learner $r = \mathcal{L}(\mathcal{D}, x)$

Example: output of feedforward neural network trained with SGD using $\mathcal{D}$ as the training set.

Rating impossibility for invariant learners

Theorem (SB, Liu, Tupper, 2022)

Consider a data set $\mathcal{D}$ and a transformation $\tau : \mathcal{X} \to \mathcal{X}$ such that

(i) $\tau(\mathcal{D}) = \mathcal{D}$ *(invariance of the data).*

Then, for any learner $\mathcal{L}$ and any input $x \in \mathcal{X}$ such that

(ii) $\mathcal{L}(\tau(\mathcal{D}), \tau(x)) = \mathcal{L}(\mathcal{D}, x)$ *(invariance of the learner),*

we have

$$\mathcal{L}(\mathcal{D}, \tau(x)) = \mathcal{L}(\mathcal{D}, x).$$

Rating impossibility for invariant learners

Theorem (SB, Liu, Tupper, 2022)

Consider a data set $\mathcal{D}$ and a transformation $\tau : \mathcal{X} \to \mathcal{X}$ such that

(i) $\tau(\mathcal{D}) = \mathcal{D}$ (*invariance of the data*).

Then, for any learner $\mathcal{L}$ and any input $x \in \mathcal{X}$ such that

(ii) $\mathcal{L}(\tau(\mathcal{D}), \tau(x)) = \mathcal{L}(\mathcal{D}, x)$ (*invariance of the learner*),

we have

$$\mathcal{L}(\mathcal{D}, \tau(x)) = \mathcal{L}(\mathcal{D}, x).$$

Proof.

$$\mathcal{L}(\mathcal{D}, \tau(x)) = \mathcal{L}(\tau(\mathcal{D}), \tau(x)) = \mathcal{L}(\mathcal{D}, x).$$

□

Invariance of the learner under SGD training

Setting: $\mathcal{D} = \{(x_i, r_i)\}_{i=1}^m$, $r = \Phi(V, Wx)$, with (V, W) trainable.

Theorem (SB, Liu, Tupper, 2022)

Let $\tau : \mathcal{X} \to \mathcal{X}$ be a linear transformation represented by an orthogonal matrix T. Compute (V_k, W_k) via k iterations of SGD from randomly initialized (V_0, W_0) and with the regularized loss

$$G(V, W) = F\left((\Phi(V, Wx_i), r_i)_{i=1}^m\right) + \lambda(R(V) + \|W\|_F^2),$$

with $\lambda \geq 0$ and such that G is differentiable. Let V_0 and W_0 be independent and $W_0 T \overset{d}{=} W_0$ (equidistributed). Then, the learner

$$\mathcal{L}(\mathcal{D}, x) = \Phi(V_k, W_k x)$$

is invariant to τ in distribution (i.e., $\mathcal{L}(\mathcal{D}, x) \overset{d}{=} \mathcal{L}(\tau(\mathcal{D}), \tau(x))$).

Extensions: Adam, models $r = \Phi(V, Wx, Wy)$ (e.g., recurrent NNs)

Invariance of the data and the transformation τ

Back to our initial example, define $\begin{cases} \tau(\ell_1 \mathsf{Y}) = \ell_1 \mathsf{Z}, \\ \tau(\ell_1 \mathsf{Z}) = \ell_1 \mathsf{Y}, \\ \tau(\ell_1 \ell_2) = \ell_1 \ell_2, \quad \forall \ell_2 \notin \{\mathsf{Y}, \mathsf{Z}\}. \end{cases}$

- $\tau(\mathcal{D}) = \mathcal{D}$ (recall that $\ell_1 \mathsf{Y}, \ell_1 \mathsf{Z} \notin \mathcal{D}$).
- If we encode letters as $\{\mathsf{A}, \mathsf{B}, \ldots, \mathsf{Z}\} \to \{v_j\}_{j=1}^{26} \subset \mathbb{R}^{26}$, τ is linear.

Encoding $\{v_j\}_{j=1}^{26}$	**Property of matrix T**	**Invariance of $\mathcal{L}$?**
canonical basis (one-hot)	permutation	✓
orthogonal basis	orthogonal	✓
linearly independent	invertible	?

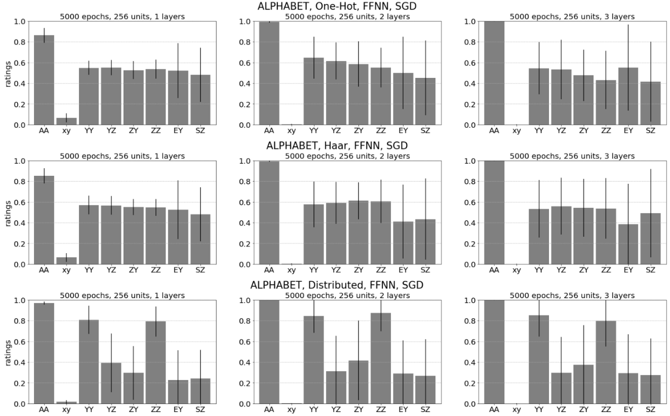
So, if $\mathcal{L}$ is invariant to τ, then $\mathcal{L}(\mathcal{D}, \tau(x)) \overset{d}{=} \mathcal{L}(\mathcal{D}, x)$.

In particular, if $x = \mathsf{YZ}$, then $\tau(x) = \mathsf{YY}$ and $\mathcal{L}(\mathcal{D}, \mathsf{YY}) \overset{d}{=} \mathcal{L}(\mathcal{D}, \mathsf{YZ})$.

$\Rightarrow$ **The learner $\mathcal{L}$ is unable to generalize outside the training set.**

Numerical experiment: Two-letter words

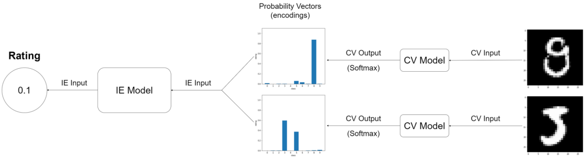
Learners: (left to right) feedforward NN (depth = 1, 2, 3)
Encodings: (top to bottom) one-hot, orthogonal, distributed.



Bar 1-2: words in the training set.
Bar 3-8: words outside the training set (contain Y or Z).

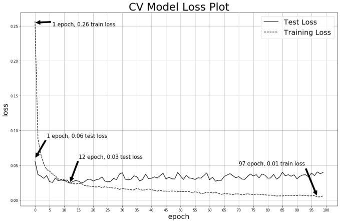
Handwritten digits (MNIST): Setup

Task: Classify pairs of images corresponding to palindromic numbers.



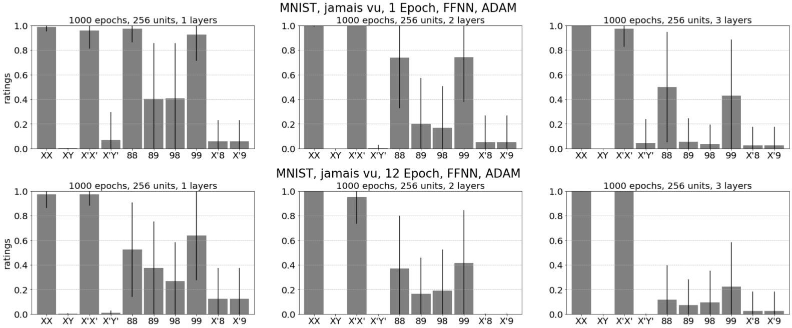
Setup:

- Computer Vision (CV) model (trained on all digits)
- Identity Effect (IE) model (trained on digits from 0 to 7)



Handwritten digits (MNIST): Results

Learners: (left to right) IE feedforward NN (depth = 1, 2, 3)
Encodings: (top to bottom) undertrained CV, overtrained CV



Bars 1-2: images in the training set
Bars 3-4: unseen images (but already seen digits)
Bars 6-10: digits outside the training set (contain 8 or 9)

II. Practical existence theorems

Motivation

We consider the problem of approximating a multivariate function

$$x \mapsto f(x), \quad x \in \mathbb{R}^d,$$

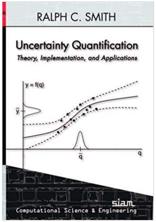
from pointwise samples $f(x_1), \ldots, f(x_m)$.

f typically arises from a parametric model describing a physical process

Parameters x	$\rightarrow$	Parametric model (e.g., parametric PDE) $\mathcal{M}_x[u]=0$ (computing $u = u_x$ is expensive)	$\rightarrow$	Quantity of interest $f(x)=\mathcal{Q}[u]$

Key tasks: surrogate modelling, uncertainty quantification.

Applications: weather and climate, epidemiology, subsurface hydrology, nuclear reactor design, biological models, ...



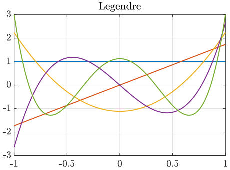
Suggested reading
[Smith, 2014]

Orthogonal polynomials and sparse approximation

Let $f : \mathcal{U} \to \mathbb{C}$, where $\mathcal{U} = [-1,1]^d$, and

$$\Psi_{\boldsymbol{\nu}} = \Psi^{1D}_{\nu_1} \otimes \cdots \otimes \Psi^{1D}_{\nu_d},$$

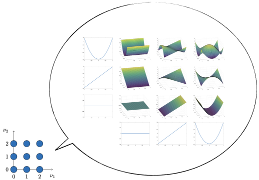
where $\{\Psi^{1D}_{\nu}\}_{\nu \in \mathbb{N}_0}$ are 1D orthogonal polynomials on $[-1,1]$ (e.g., Legendre).



$\{\Psi_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu} \in \mathbb{N}_0^d}$ orthonormal basis of $L^2(\mathcal{U})$.

For any $f \in L^2(\mathcal{U})$ we have the expansion

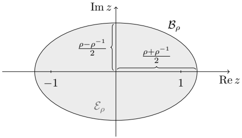
$$f = \sum_{\boldsymbol{\nu} \in \mathbb{N}_0^d} c_{\boldsymbol{\nu}} \Psi_{\boldsymbol{\nu}}, \quad c_{\boldsymbol{\nu}} = \int_{\mathcal{U}} f(x) \Psi_{\boldsymbol{\nu}}(x) \, \mathrm{d}x.$$



Goal: Compute a sparse approximation

$$f \approx \hat{f} = \sum_{\boldsymbol{\nu} \in \mathbb{N}_0^d} \hat{c}_{\boldsymbol{\nu}} \Psi_{\boldsymbol{\nu}}, \quad \|\hat{c}\|_0 = \#\{\hat{c}_{\boldsymbol{\nu}} \neq 0\} \text{ “small”}.$$

Smoothness $\Rightarrow$ Exponential best s-term decay



The Bernstein ellipse $\mathcal{E}_\rho$.

Assume that f is **holomorphic** (or **analytic**) in a **Bernstein polyellipse** $\mathcal{E}_{\boldsymbol{\rho}}$, where $\mathcal{E}_{\boldsymbol{\rho}} = \mathcal{E}_{\rho_1} \times \cdots \times \mathcal{E}_{\rho_d} \subset \mathbb{C}^d$,. Then, for $s \geq \bar{s}$,

$$\underbrace{\inf_{\|\hat{c}\|_0 \leq s} \left\| f - \sum_{\boldsymbol{\nu} \in \mathbb{N}_0^d} \hat{c}_{\boldsymbol{\nu}} \Psi_{\boldsymbol{\nu}} \right\|_{L^2}}_{\text{best } s\text{-term approx. error}} \lesssim \|f\|_{L^\infty(\mathcal{E}_{\boldsymbol{\rho}})} \cdot \exp(-\gamma s^{1/d}), \quad \gamma = \gamma(d, \boldsymbol{\rho}).$$

This holds for a large class of parametric models: diffusion equation, harmonic oscillator, heat equation, parametrized domain, ...

[Cohen, DeVore, Schwab, 2010-2011], [Chkifa, Cohen, Schwab, 2015], [Beck, Nobile, Tamellini, Tempone, 2015], [Cohen, DeVore, 2015], [Tran, Webster, Zhang, 2017]

Known vs. unknown anisotropy

Issue: **Anisotropy** of f, i.e., how smooth f is in each variable, might be unknown (can be measured by $\boldsymbol{\rho}$).

Sparse polynomial approximation methods:

Known anisotropy

- **Interpolation** via *sparse grids* [Zenger, 1991],[Bungartz, Griebel, 2004]
- **Quadrature methods** (approximate $c_{\boldsymbol{\nu}} = \int_{\mathcal{U}} f(x)\Psi_{\boldsymbol{\nu}}(x)\,\mathrm{d}x$)
- **Least-squares approximation**

$$\min_{p \in \mathrm{Span}\{\Psi_\nu\}_{\nu \in S}} \frac{1}{m} \sum_{i=1}^{m} |p(\boldsymbol{y}_i) - f(\boldsymbol{y}_i)|^2$$

Unknown anisotropy

- **Greedy (adaptive) methods**
- **Compressed sensing** $\leftarrow$ $\boxed{\text{This talk}}$
- **Deep learning** $\leftarrow$ $\boxed{\text{This talk}}$

High-dimensional approximation via compressed sensing

- Consider a "large enough" ambient set $\Lambda \subset \mathbb{N}_0^d$, $|\Lambda| = N$ and let

$$f_\Lambda = \sum_{\nu \in \Lambda} c_\nu \Psi_\nu$$

- Collect Monte Carlo (MC) samples, i.e. $x_1, \ldots, x_m \in \mathcal{U}$ i.i.d. uniform samples, where $m \ll N$.
- Let $A = (\frac{1}{\sqrt{m}} \Psi_{\nu_j}(x_i))_{i,j=1}^{m,N} \in \mathbb{C}^{m \times N}$, $b = (\frac{1}{\sqrt{m}} f(x_i))_{i=1}^m \in \mathbb{C}^m$.
- Obtain underdetermined linear system $b = Ac_\Lambda + e$, where

$$\underbrace{c_\Lambda = (c_{\nu_j})_{j=1}^N}_{\text{coefficients of } f_\Lambda}, \quad e = \frac{1}{\sqrt{m}}(\underbrace{f(x_i) - f_\Lambda(x_i)}_{\substack{\text{truncation}\\ \text{error}}} + \underbrace{n_i}_{\substack{\text{more sources}\\ \text{of error}}})_{i=1}^m$$

- Let $u_\nu = \|\Psi_\nu\|_{L^\infty}$, solve the Square Root LASSO

$$\hat{c} \in \arg\min_{z \in \mathbb{C}^N} \|Az - b\|_2 + \lambda \|z\|_{1,u}, \quad \hat{f} = \sum_{\nu \in \Lambda} \hat{c}_\nu \Psi_\nu.$$

[Doostan, Owhadi, 2011], [Mathelin, Gallivan, 2012], [Yang, Karniadakis, 2013], [Rauhut, Ward, 2016], [Adcock, 2017], [Chkifa, Dexter, Tran, Webster, 2018], [Adcock, SB, Webster, 2018]

Convergence rates for compressed sensing

Theorem [Adcock, SB, Webster, 2022], [Adcock, SB, Dexter, Moraga, 2021]

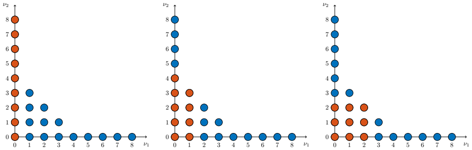
Let f be holomorphic in $\mathcal{E}_\rho$ and set $\widetilde{m} = cm/(\log^3(m)\log(d))$. Let

$$\Lambda := \Lambda^{\mathrm{HC}}_{d,s-1} = \left\{ \boldsymbol{\nu} \in \mathbb{N}_0^d : \prod_{k=1}^{d} (\nu_k + 1) \leq s \right\}$$

be the hyperbolic cross index set of order $s = \lceil \widetilde{m}^{1/2} \rceil$. Then,

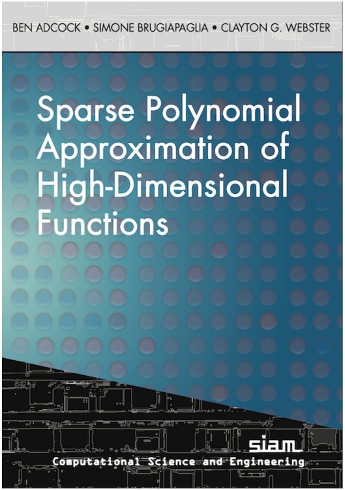
$$\|f - \hat{f}\|_{L^2} \lesssim \|f\|_{L^\infty(\mathcal{E}_\rho)} \cdot \exp(-\gamma \widetilde{m}^{1/(2d)}) + \frac{1}{\sqrt{m}} \|n\|_2,$$

with high probability.



Key fact: $\Lambda^{\mathrm{HC}}_{d,s-1}$ contains all "lower sets" of cardinality s.

If you want to know more...



http://sparse-hd-book.ca

DNN Approximation Theory

Generalizations of **universal approximation theorems** proposed in the 80s by Chibenko, Hornik et al., show that DNNs can efficiently approximate functions from a wide variety of classes:

- E.g. H^k functions, piecewise smooth functions, bandlimited functions, Barron functions, cartoon-like functions,...

  Review paper [Elbrächter, Perekrestenko, Grohs, Bölcskei, 2021]

For holomorphic functions there exist DNNs (of moderate size and depth) that achieve the same error bounds as the best s-term polynomial approximation. [Opschoor, Schwab, Zech, 2019], [Daws, Webster, 2020] [Adcock, SB, Dexter, Moraga, 2021]

Key questions

1. Can DNNs with suitable approximation properties be obtained via training? How much data do we need?
2. How does DNN-based approximation compare with polynomial approximation via CS?

Practical existence theorem for DNNs

Theorem [Adcock, SB, Dexter, Moraga, 2021]

Let f be holomorphic in $\mathcal{E}_{\boldsymbol{\rho}}$, $\{x_i\}_{i=1}^m$ i.i.d. uniform samples from $\mathcal{U}$ and define $\widetilde{m} := cm/(\log^3(m)\log(d))$. Then, there exist

- a class of ReLU DNNs $\mathcal{N}$ whose depth, # of trainable parameters, and # of nonzero parameters, are at most polynomial in $\widetilde{m}$;
- a regularization functional $\mathcal{R} : \mathcal{N} \to [0, \infty)$ equal to a certain norm of the trainable parameters

such that any minimizer

$$\hat{\Phi} \in \arg\min_{\Phi \in \mathcal{N}} \left(\frac{1}{m} \sum_{i=1}^{m} |\Phi(x_i) - f(x_i)|^2 \right)^{1/2} + \lambda \mathcal{R}(\Phi),$$

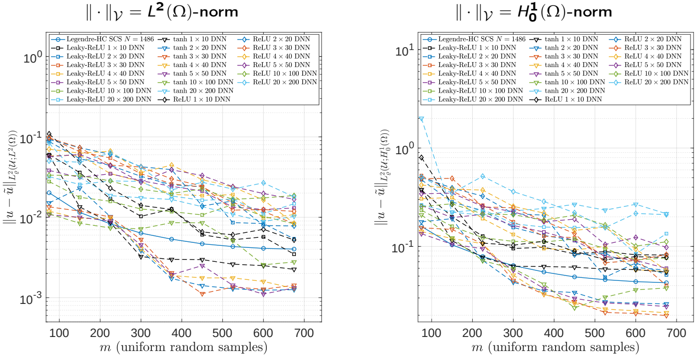
satisfies the same exponential convergence rates in $\widetilde{m}$ as those for sparse polynomial approximation via CS with high probability.

Extensions: Hilbert- and Banach-valued settings [Adcock, SB, Dexter, Moraga, 2022].

Numerics: parametric diffusion equation

[Adcock, SB, Dexter, Moraga, 2021]

Parametric PDE: $d = 30$ dimensional parametric diffusion equation on $\Omega = [0,1]^2$ with "layered" spatial dependence (based on benchmark from [Nobile, Tempone, Webster, 2008]).



Take home: By careful tuning of the architecture, DNNs can achieve the similar or better performance than CS.

Epilogue

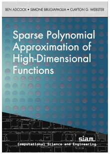
- Inspired by Smale's 18th problem, we have illustrated how ideas from geometry, approximation, probability, and optimization can help provide new insights on the mathematical foundations of deep learning.
- In particular, we have seen results that identify limitations and potential of deep learning in different contexts: identity effect classification and high-dimensional approximation.
- We have only scratched the surface, and much more work remains to be done!

The roads not taken...

- Rating impossibility with noisy encodings
  with **Paul Glickman** (Concordia)
- Rating impossibility for Graph Neural Networks (GNNs)
  with **Alessio D'Inverno** (University of Siena & MILA) and Mirco Ravanelli (Concordia & MILA)
- Numerical approximation of high-dimensional PDEs via compressed sensing and deep learning
  with Nick Dexter (FSU) and **Weiqi Wang** (Concordia)
- Analysis of compressive sensing with deep generative priors
  with **Aaron Berk** (McGill), Babhru Joshi (UBC), Yaniv Plan (UBC), Matthew Scott (UBC), and Özgür Yilmaz (UBC)
- Sparse recovery and deep algorithm unrolling
  with **Sina M.-Taheri** (Concordia)

Thank you!

Book



B. Adcock, SB, and C. Webster, **Sparse Polynomial Approximation of High-dimensional Functions**, SIAM, 2022

`www.sparse-hd-book.ca`

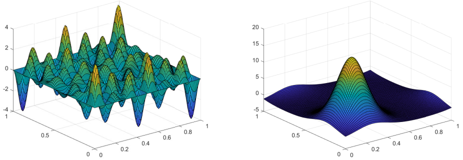
Papers

- SB, M. Liu, and P. Tupper, **Invariance, encodings, and generalization: learning identity effects with neural networks.** Neural Computation, 34 (8), pp. 1756-1789, 2022
- B. Adcock, SB, N. Dexter, and S. Moraga, **Deep neural networks are effective at learning high-dimensional Hilbert-valued functions from limited data**, Proceedings of Machine Learning Research (PMLR), MSML21 2021

Backup slides

From function approximation to high-dimensional PDEs

Sparsity and Monte Carlo sampling can help solve PDEs on high-dimensional domains, via **compressive spectral Fourier collocation**. [SB, Wang, 2022]



- Under suitable sufficient conditions on diffusion coefficient the curse of dimensionality can be lessened in the number of collocation points. Theory is based on random sampling in Bounded Riesz systems [SB, Dirksen, Jung, Rauhut, 2021]
- Practical existence theorems for DL-based high-dimensional PDE solvers? (Physics Informed NNs [Lagaris, Likas, Fotiadis, 1998], [Karniadakis et al., 2021])

Compressed sensing and deep generative models

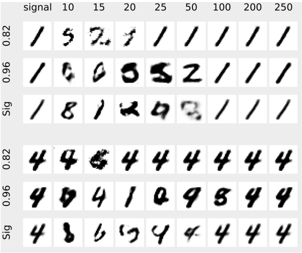
Compressed sensing can be used to recover signals in the range of a deep generative model [Bora, Jalal, Price, Dimakis, 2017]

Goal: Recover $x = G(z) \in \mathbb{R}^N$ from $m \ll N$ noisy linear measurements $y = Ax + e$, where $G : \mathbb{R}^k \to \mathbb{R}^N$ is a neural network of depth D.

In [Berk, SB, Joshi, Plan, Scott, Yilmaz, 2022] we provide the first recovery guarantees for generative compressed sensing with subsampled isometries based on a coherence parameter α. We prove that

$$m \gtrsim kDn\alpha^2$$

measurements are sufficient for accurate and stable recovery. Typical coherence (random weights) is $\alpha = O(\sqrt{kD/n})$



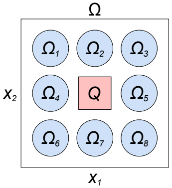
Lower coherence leads to better recovery.

Example: Parametric diffusion equation

Physical variables: $\boldsymbol{z} \in \Omega = (0,1)^2$
Subdomains: $\Omega_k \subset \Omega$ (circles) , $Q \subset \Omega$ (square)
Parameters: $x \in [-1,1]^8$

Parametric PDE: For any $x \in [-1,1]^8$, find solution $u(\cdot, x)$ to

$$\begin{cases} -\nabla_{\boldsymbol{z}} \cdot (a(\boldsymbol{z}, x) \nabla_{\boldsymbol{z}} u(\boldsymbol{z}, x)) = 1_Q(\boldsymbol{z}), & \boldsymbol{z} \in \Omega, \\ u(\boldsymbol{z}, x) = 0, & \boldsymbol{z} \in \partial\Omega, \end{cases}$$



where $a(\boldsymbol{z}, x) = \epsilon + \sum_{k=1}^{d} c_k(x_k) 1_{\Omega_k}(\boldsymbol{z}) \geq \epsilon > 0.$

Parametric solution map: $x \mapsto u(\cdot, x)$

Quantities of interest:

$$f(x) = \int_\Omega u(\boldsymbol{z}, x)\, \mathrm{d}\boldsymbol{z}, \quad f(x) = u(\boldsymbol{z}_0, x), \quad \ldots$$

Proof sketch (Practical existence theorem)

1. Define the class of DNNs

$$\mathcal{N} = \{\Phi : \mathbb{R}^d \to \mathbb{R} : \Phi(x) = z^T \Phi_{\Lambda,\delta}(x),\ z \in \mathbb{R}^N\}$$

   where

   - z are trainable parameters
   - $\Phi_{\Lambda,\delta} = (\Phi_{\boldsymbol{\nu},\delta})_{\boldsymbol{\nu}\in\Lambda}$ is a ReLU network (with explicit depth and width bounds) that approximates Legendre polynomials $\Psi_{\boldsymbol{\nu}}$ s.t. $\|\Psi_{\boldsymbol{\nu}} - \Phi_{\boldsymbol{\nu},\delta}\|_{L^\infty(\mathcal{U})} \leq \delta$ [Opschoor, Schwab, Zech, 2019]

2. The DNN training program can be interpreted as a SR-LASSO program. In particular,

$$\hat{c} \in \arg\min_{z\in\mathbb{C}^N} \|A'z - b\|_2 + \lambda\|z\|_1,$$

   where $A' = (\frac{1}{\sqrt{m}}\Phi_{\boldsymbol{\nu}_j,\delta}(x_i))_{ij} \approx A$, the CS matrix, if and only if

$$\hat{\Phi} = \hat{c}^T \Phi_{\Lambda,\delta}(x),$$

   is a minimizer to the training program.

3. Now, use tools from sparse high-dimensional polynomial approximation via CS.